BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 12 2006, pages 1431–1436
doi:10.1093/bioinformatics/btl112

Genome analysis

AuberGene—a sensitive genome alignment tool

Radek Szklarczyk and Jaap Heringa*

Centre for Integrative Bioinformatics VU (IBIVU), Faculty of Sciences and Faculty of Earth and Life Sciences, Vrije Universiteit, De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands

Received on December 15, 2005; revised on March 20, 2006; accepted on March 21, 2006
Advance Access publication April 10, 2006
Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: The accumulation of genome sequences will only accelerate in the coming years. We aim to use this abundance of data to improve the quality of genomic alignments and devise a method which is capable of detecting regions evolving under weak or no evolutionary constraints.

Results: We describe a genome alignment program *AuberGene,* which explores the idea of transitivity of local alignments. Assessment of the program was done based on a 2 Mbp genomic region containing the CFTR gene of 13 species. In this region, we can identify 53% of human sequence sharing common ancestry with mouse, as compared with 44% found using the usual pairwise alignment. Between human and tetraodon 93 orthologous exons are found, as compared with 77 detected by the pairwise human-tetraodon comparison.

AuberGene allows the user to (1) identify distant, previously undetected, conserved orthogonal regions such as ORFs or regulatory regions; (2) identify neutrally evolving regions in related species which are often overlooked by other alignment programs; (3) recognize false orthologous genomic regions. The increased sensitivity of the method is not obtained at the cost of reduced specificity. Our results suggest that, over the CFTR region, human shares 10% more sequence with mouse than previously thought (∼50%, instead of 40% found with the pairwise alignment).

Availability: The source code and tracks for UCSC Genome Browser generated with the program are available from http://www.ibivu.cs.vu.nl/programs/aubergenewww.

Contact: heringa@cs.vu.nl

1 INTRODUCTION

The rapid rate of accumulation of entire genomic sequences has fueled the development of the comparative genomics field. With more data at hand the results of analysis are more sensitive to weaker conservation signals and statistically sound at the same time. However, with even more genomic sequences nearing completion, the need for fast, reliable and automatic tools for alignment, with the emphasis on specificity and permitting alignment of neutrally evolving regions, is growing (Schwartz *et al*., 2003). In this paper we use intermediate sequences in order to be able to delineate significantly divergent regions of homologous sequences. This idea has been applied previously in the area of sequence analysis, in such diverse tasks as homology detection among proteins (Park *et al*., 1997), multiple alignment (Morgenstern *et al*., 1998; Notredame *et al*., 2000; Ye and Huang, 2005), repeat detection (Szklarczyk and Heringa, 2004) and identification of weak-signal protein sequence motifs (Heger *et al*., 2004). It has been especially successful in inferring distant relationships where homology cannot be detected by simple, direct pairwise comparison.

Here we employ transitivity for the analysis of genomic sequences, concentrating on pairwise local alignments. Transitivity enables us to align regions that are difficult to align (in particular regions that are distant) by identifying orthologous fragments such as exons or regulatory elements. The increase in quality of pairwise alignment extends beyond the ability to cover longer evolutionary distances or to find more matches between sequences. It also allows the identification of false-positive, non-homologous alignments which can be corrected based on the new information provided by intermediate sequences. Alignments that are confirmed by matches involving different intermediate genomes suggest the functional importance of such genomic fragments [as it is observed that regions conserved in multiple species often correlate with functional elements (Kellis *et al*., 2003)].

2 ALGORITHM

We start the procedure by calculating all-against-all pairwise genome alignments. This is done by running genome alignment software such as BLASTZ (Schwartz *et al*., 2003). We refer to these alignments as direct alignments.

We note that an alignment between a pair of sequences (e.g. human and fugu, Fig. 1a) can also be produced using a third sequence (e.g. chicken) as a result of combining the human–chicken and chicken–fugu alignments. The alignment obtained in this transitive step comprises matches between residues of human and fugu which were matched with the same chicken residue (Fig. 1b). This process results in the transitive pairwise alignment human–chicken–fugu (Notredame *et al*., 2000; Szklarczyk and Heringa, 2004). In this case, even though three sequences and two alignments were used for the process, the result is a pairwise alignment.

Using a set of N genomes, $N - 2$ transitive alignments can be formed for a given pair of genome sequences. To fully see the benefits of the whole set of alignments which we have at hand at this point, we merge them into a single alignment (Fig. 2). When merging, for each match we keep track of the number of transitive alignments which include this match—a match between certain residues can be indicated by many transitive alignments independently. The result of the merging process, the collective alignment, contains matches with weights ranging from 1 (when there was only

*To whom correspondence should be addressed.

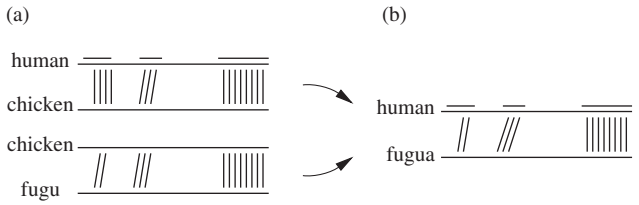

Fig. 1. Calculating the transitive human–chicken–fugu alignment. (**a**) Two direct alignments: human–chicken and chicken–fugu and (**b**) the transitive human–fugu alignment, based on the two alignments with chicken. Only residues that match the same chicken residue are matched in the transitive alignment.

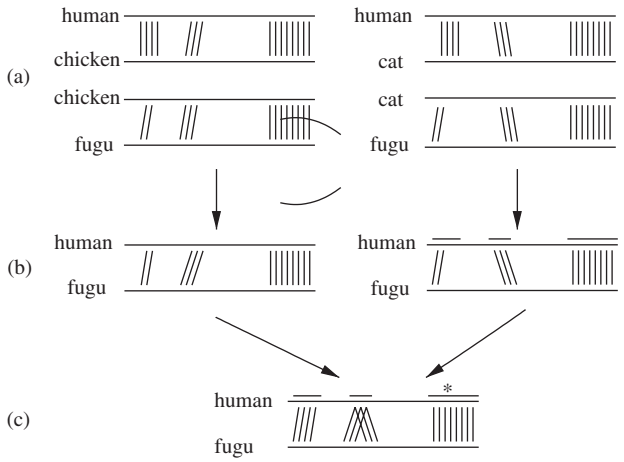

Fig. 2. Calculating the collective human–fugu alignment using two intermediate sequences. (**a**) Direct alignments with chicken and cat, (**b**) transitive human–fugu alignments and (**c**) the collective alignment: created by combining the two transitive alignments. The middle region is matched inconsistently, and the region denoted with a star is confirmed both directly and transitively.

one transitive alignment with this match) to $N - 2$ (in which the whole set of transitive alignments supports the match). The weight reflects our confidence in the match, normally expressed as the alignment score. The collective alignment allows us to (1) increase the coverage of the sequence (sensitivity) and thereby extend homology detection to regions not represented in the direct alignment; (2) confirm matches (gaining confidence) in the direct alignment for those residues consistently matched in the transitive alignments and (3) contradict (or reduce the importance of) direct matches due to inconsistent matching in the various transitive alignments. The latter can be used to decide whether the direct alignment indicates a true homology, or results from an artifact of the method used to produce it (specificity).

Alignments are internally represented as gapless high-scoring segments. This approach is favored over storing alignments as a set of matches, which could, for long genomic alignments, lead to the increase of space/time complexity. Our algorithm constructs the collective alignment in three steps.

Step 1. *Segment decomposition*. Gapless segments which overlap partially with other segments are parsed, such that only fully overlapping or non-overlapping segments are produced. Initially,

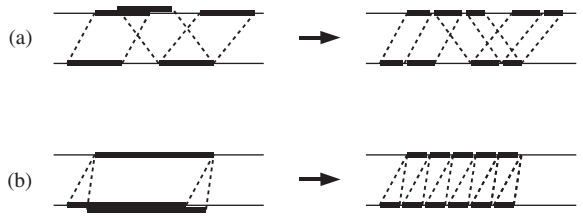

Fig. 3. Breaking up partially overlapping segments into non- or fully overlapping ones. (**a**) A typical case and (**b**) two partially overlapping segments lead to a large number of fully overlapping segments.

we merge all the transitive alignments and the direct one together. Therefore some of the gapless segments may begin or end inside another segment, and it is likely that a number of residues from one sequence will have multiple matches with residues in the other sequence (Fig. 3a). We adopt a depth-first search strategy to parse the segments. First, we mark all residues within each of the segments that are at a position where another segment begins or ends. For each of the thus marked residues we then follow its links to matched positions in the other sequence and mark those as well. We recursively follow the links going out of these newly marked residues and mark the corresponding residues until no new unmarked positions can be found anymore. Upon termination of the procedure, we split each of the initial segments at marked positions to create new uninterrupted segments, that is, with marks appearing at the beginning or end only. This scenario results in decomposition of partially overlapping segments into fully overlapping or non-overlapping segments. Note that two partially overlapping segments can lead to an arbitrary number of fully overlapping ones (Fig. 3b).

Step 2. *Constructing a weighted bipartite graph from transitive alignments*. For this parsed genome alignment, we create weighted edges between each segment in either genome, where the weight corresponds to the number of alignments (direct and transitive) that support the alignment of the segment pair. The higher the integer weight, the more transitive alignments support the segment matching, which is a strong indication of orthology.

Step 3. *Generating the collective alignment*. After a genome alignment is converted to a weighted bipartite graph, the final alignment is determined by running the Hopcroft–Karp bipartite max-cardinality matching algorithm (Hopcroft and Karp, 1973) (implemented after Cormen *et al*., 2001). This algorithm selects the maximal subset of the edges that do not share a common node. This corresponds to choosing segments of the merged alignment in such a way that no residue is covered by more than one match, while at the same time the number of matches and thereby the coverage of the alignment is maximized. Our method provides an option to skip this step of the algorithm that can be used if the user wishes to retain alignments with multiple matched segments.

3 RESULTS

3.1 Sensitivity assessment

In our analysis we used the data from the greater CFTR region containing the gene encoding the cystic fibrosis transmembrane conductance regulator and nine other genes [1.8 Mbp on human chromosome 7, around 1.4 Mbp for the other mammals and around 0.3 Mbp for the non-mammalian species (Thomas *et al*., 2003)]. The

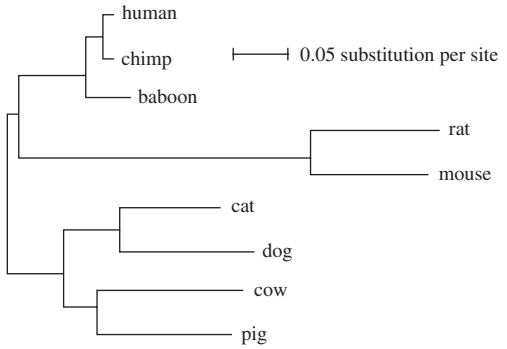

Fig. 4. Phylogenetic tree for the mammalian species used in this study, adapted from Thomas *et al*. (2003).

region has been sequenced in 13 different species: baboon, cat, chicken, chimp, cow, dog, fugu, human, mouse, pig, rat, tetraodon (tetra in figures) and zebrafish (zfish). The phylogenetic tree for the mammalian organisms is shown in Figure 4 where it is apparent that the rat and mouse genomes evolve the fastest.The sequences were aligned all-against-all (Thomas *et al*., 2003) using BLASTZ (Schwartz *et al*., 2003). We found that for most organisms, transitivity leads to a significantly higher fraction of sequence aligned with human. The increase of the length of the alignment (later we will argue that these are not spurious matches) varies between 1.5 kbp (zebrafish) and 140 kbp (rat and mouse), i.e. between 0.5 and 10% of the human sequence in this region (Fig. 5a). Interestingly, we observed that most new putative orthologous nucleotides are found between human and rodents. In general, the fraction of aligned sequence decreases with evolutionary distance with the exception of rat and mouse. Rodents, even though considered evolutionary closer to human than other non-primate mammals considered here (Murphy *et al*., 2001, Fig. 4), have a lower fraction of sequence aligned with human than cat, cow, dog and pig (Thomas *et al*., 2003, Fig. 5a). The gap between rodents and other mammals is partially bridged by the use of transitive alignments, suggesting that the method performs very effectively at finding homologous fragments of sequence for genomes suspected to evolve at a particularly high rate (Mouse Genome Sequencing Consortim, 2002, Fig. 4). We found that new matches occur more frequently in intronic regions than in regions between genes. For the sequences analyzed, the transitive matches are found in introns almost twice as often (after discounting for the length difference) as the (in principle) untranscribed regions.

Overall, ∼44% of the human greater CFTR region is covered by the direct alignment with mouse, which is in close agreement with the whole genome alignment covering 40% of the 2.9 Gb human sequence (Mouse Genome Sequencing Consortium, 2002). In the paper announcing the publication of the mouse genome authors report alignment of the most of orthologous sequence, stating that the rest is likely to have been deleted in one or both genomes. Our method increases the length of the aligned sequence by 10 percentage points, to 53% in the CFTR region (Fig. 5a).

In order to explain which species predominantly contribute to the increased coverage of human–mouse alignment, new residues covered with matches were counted separately for each of the intermediate sequence (Fig. 6). The second rodent, rat, is the most helpful in enriching the human–mouse alignment, followed by

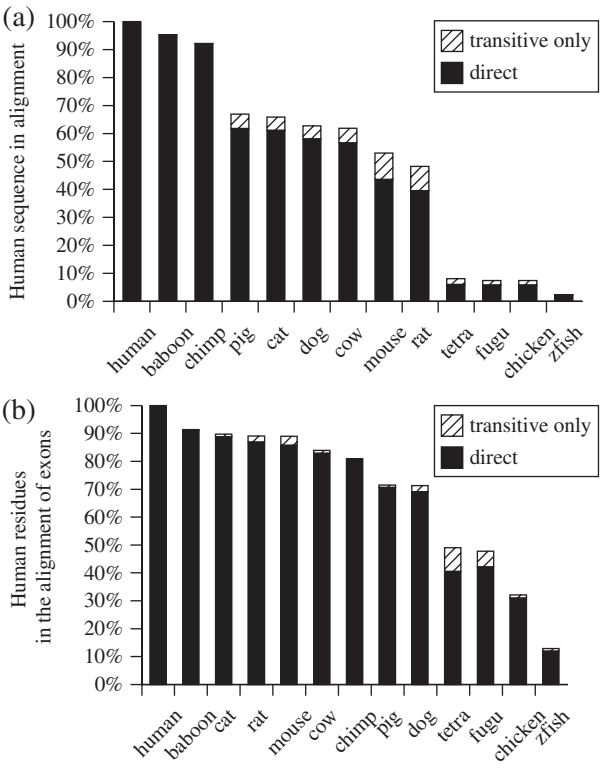

Fig. 5. Coverage of human sequence using direct (black) and transitive alignments (hatched) of the greater CFTR region. Each bar represents the number of nucleotides aligned between human and other species (horizontal axis). (**a**) Percentage coverage over the greater CFTR region and (**b**) percentage coverage over 125 human exons in this region (33 kbp in total).

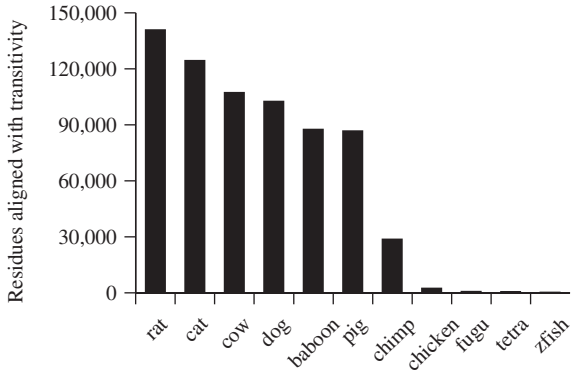

Fig. 6. Number of orthologous residues in human–mouse alignment determined by using individual intermediate genomic sequences. The total nucleotide coverage of the human genomic sequence in the human–mouse collective alignment exceeds 800 kbp.

the other mammals. We observed about one-third increase in alignment length for the species most divergent from human (Fig. 7), a consistent increase in relative alignment size for rodents (>20%) and 8–9% for other non-primate mammals investigated.

In exonic regions, the most notable increase in sensitivity is observed, not surprisingly, for very distant species: fugu and

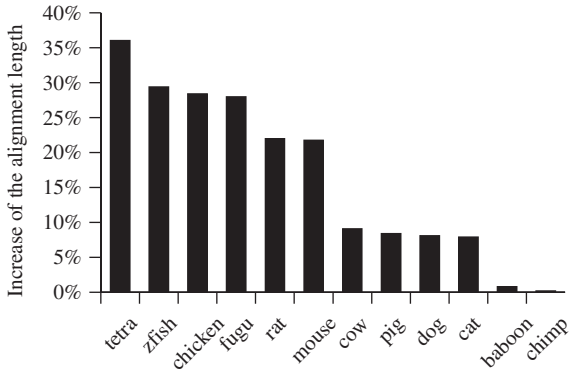

Fig. 7. Increase in length of collective alignments relative to the length of the corresponding direct alignments.

tetraodon (Fig. 5b). No new matches for exons were found using primates. This is most likely due to the fact that all orthologous exons were already indicated by the pairwise alignment, and the only reason that the alignment does not cover all of the coding sequence is exon-loss in one of the lineages. There is a consistent increase in the coverage of human exons in alignments with non-primate mammals, with the increase being highest for rodents, as expected from overall coverage of the sequence (Fig. 5a).

To verify that transitive alignments are able to cover larger evolutionary distances (which would mean that the method is sensitive for homology of distant, but functional regions), we searched for orthologous exons of the greater CFTR region between two very divergent sequences: human and tetraodon. Given such a distant sequence comparison, only 23% of the the whole collective alignment is outside of boundaries of human exons. Despite that, out of 125 human exons in the region [as annotated in GenBank (Benson *et al*., 2004), release February 2004], the pairwise alignment finds 77 orthologous exons. The collective alignment with 11 intermediate genomes increases the number of detected orthologous human exons to 93 (Table 1).

3.2 Specificity assessment

To assess specificity we reversed (without complementing) all genomic sequences, and used them as an intermediate for constructing the collective alignment of human and mouse sequences (therefore creating transitive alignments such as human–dog^R–mouse). In this case, the reversed sequence acts as a randomized sequence but preserves the local composition of nucleotides. Alignment with such a reversed genome approximates (although slightly underestimating) the quantity of spurious matches (for details see Schwartz *et al*., 2003). Merging all such transitive alignments we created the collective alignment which covered only 0.03% of the human sequence. This demonstrates the high specificity of the method suggesting that the collective alignment we build is not likely to contain many spurious matches.

Knowing that our method does not lead to many false matches, we verified the direct human–mouse alignment using the transitive alignments. The collective alignment matched 11% of human nucleotides differently than in the direct alignment (this translates to 200 kbp in the region considered, covering the human genome evenly). For 1% of the direct human–mouse alignment, we find

Table 1. Additional 16 orthologous human exons detected in the collective human–tetraodon alignment.

Gene name	Exon position	
	start	end
CAV2	329316	329465
MET	570386	570559
	587172	587308
	592584	592802
	599179	599325
	604415	604645
WNT2	1152441	1152823
GASZ	1251771	1251893
CFTR	1309629	1309681
	1338568	1338676
	1424464	1424592
	1444147	1444247
CORTBP2	1549038	1549170
	1557623	1557801
	1575365	1575464
	1609975	1610125

The exon positions are given relative to the start of the greater CFTR region.

indications to correct and re-align human residues (where at least three transitive alignments consistently suggest a different match).

As an example, we show the conservation of the sequence in the neighborhood of the last exon in the CAV1 gene is shown in Figure 8. In contrast to the identity levels, the coding region receives the maximal confidence weight of 12 (i.e. the region is consistently aligned with mouse for all the transitive alignments). The UTR of this gene, with the signals specifying the way RNA is to be used and the rate of poly-A shortening (Alberts *et al*., 2002) is more consistently aligned in the collective alignment. A high confidence weight of 8 is visible in the intronic region 100 nt upstream from the exon, suggesting a much greater functional role than implied by identity levels in the direct alignment alone.

Even though we use transitive alignments to extend the direct one, they should, at least partially, overlap. The overlap with the direct alignment serves both as confirmation of its validity, and as an indication that transitive alignments do not produce spurious matches (with most of the transitive alignment expected to overlap with the direct, Table 2). Not surprisingly, different intermediate sequences support the direct alignment in varying degree: starting from as low as 1% (zebrafish) up to 78% (baboon), the lower percentages resulting from short transitive alignments. However, the percentage of matches of transitive alignments overlapping with the direct one is very high, exceeding 73% for all the organisms.

The program AuberGene is available for download at http://www.ibivu.cs.vu.nl/programs/aubergenewww, together with the information how to use it. The program allows the user to

- create transitive alignments
- merge alignments, allowing for multiple matches with a single nucleotide, and assign weights corresponding to the number of overlapping matches.
- make the collective alignment (with at most one match per nucleotide)

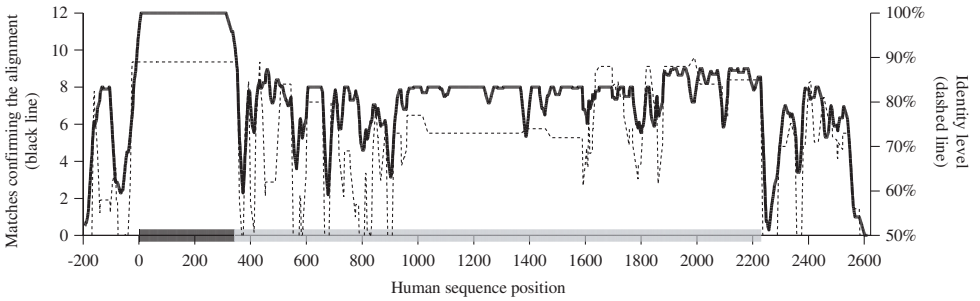

Fig. 8. Conservation of human sequence in the human–mouse alignment of the last exon of the human CAV1 gene. The number of consistent matches in the direct and transitive alignments is plotted with a thick, black line. The dashed line denotes identity levels in the direct alignment only. We marked along the horizontal axis both the exon (with a thick black line) and UTR region of the gene (grey line). Positions in the human sequence are counted starting from the last CAV1 exon.

Table 2. Overlap between direct (d) and transitive (t) human–mouse alignments

Intermediate species	$\lvert d \cap t \rvert / \lvert d \rvert$ %	$\lvert d \cap t \rvert / \lvert t \rvert$ %
baboon	78	85
chimp	71	94
rat	70	76
cat	51	73
cow	45	73
dog	42	73
pig	38	74
chicken	2	83
fugu	2	94
tetra	2	94
zfish	1	92

The first column lists organisms used as an intermediate to construct the transitive human–mouse alignment t. The second column indicates the fraction of corresponding matches in the direct alignment: the matches identical in the two alignments are denoted by $d \cap t$ and $\lvert d \rvert$ denotes the number of matches in the direct alignment. The third column gives the fraction of corresponding matches in the transitive alignment.

- report conflicting matches between alignments
- filter the alignment, restricting it to a specific region or a match weight

AuberGene creates a collective human–mouse alignment of almost 2 Mbp region for 11 intermediate genomes in <4 min on a 1.6 GHz Pentium III processor, using 100 MB of memory at peak.

Memory complexity of the program is linear and depends on the number of gapless segments in alignments (this tends to be a much lower number than the sequence length). The upper bounds on time complexity are set by a sorting procedure of n gapless segments in $O(n \log n)$ and by the Hopcroft–Karp bipartite max-cardinality matching algorithm. The latter runs in $O(\sqrt{V}E)$ time, where V and E are the numbers of vertices and edges in the bipartite graph, respectively. Because we run this procedure for each connected graph and these graphs are generally small, there is no significant impact on the program performance.

On the web page we made available so-called 'custom tracks' visualizing the greater CFTR region using UCSC Genome Browser (Karolchik *et al*. 2003, http://genome.ucsc.edu). These tracks show the coverage of human sequence with matches in human–mouse alignments. Coverage for both collective alignment (where next to 11 transitive the direct alignment is incorporated) and, for comparison, direct alignment tracks is visible. Another track represents fragments of sequence (visible as features in the browser) where the direct alignment is inconsistent with the collective one. Each feature is annotated with a number ranging from 1 to 12, i.e. the number of transitive alignments supporting the match.

4 DISCUSSION

In this work we show that the concept of transitivity can be applied effectively to genome alignments, which provides the opportunity to align genomes with increased sensitivity. Because the focus of genome alignment methods is on determining orthology, transitivity helps to find functional regions under selection pressure in distant species as well as neutrally evolving regions in closer related species.

Having a method to extend and validate alignments we do not have to put so much emphasis on the scoring method (Vingron and Waterman, 1994) and alignment strategies (Zhang *et al*., 1999). This feature is very important since genome alignment strategies, having sacrificed generality for speed, use many heuristics to rapidly process large amounts of data. Genome alignment tools are designed for efficient comparison of sequences at a certain evolutionary distance (such as between human and mouse, Miller, 2001), and are therefore suboptimal for more divergent genomes. Not only these heuristics, but also some intrinsic properties of the local alignment technique lead to potential flaws (Arslan *et al*., 2001), for example, the inclusion of an arbitrary poor internal segment in an alignment (Zhang *et al*., 1999).

When running AuberGene, the user normally does not need to pay particular attention to which sequences are used as intermediates—in fact the greater the number of sequences, the higher the coverage. Nonetheless, if intermediate sequences are included that are very closely related to one or both of the sequences considered, due to the inherent support by these sequences the weights of the original direct alignment will tend to increase. This might lead to reduced additional information and will make the program run longer.

Often, genomic alignments do not have a 1-to-1 relationship at the residue level, either due to segmental duplications or spurious hits, leaving alignment tools with no option but to provide output with multiple matches. Indeed, experience with genome alignments

suggests that the initial alignment should be fairly inclusive, and decisions about processing it should be left to downstream tools (Schwartz *et al*., 2003). We have presented such a downstream tool here, and have shown that by including additional information from other species we can improve sensitivity considerably and produce an alignment that is more accurate and less ambiguous.

ACKNOWLEDGEMENTS

The authors would like to thank Evert Wattel, for advice on graph-theory problems, and David Eppstein, for the python implementation of the Hopcroft–Karp bipartite max-cardinality matching algorithm (implemented in Python after Cormen *et al*., 2001). We also thank Nicola Armstrong and Wojciech Makalowski, for valuable suggestions leading to improvement of the manuscript. The research was funded by the Vrije Universiteit of Amsterdam.

Conflict of Interest: none declared.

REFERENCES

Alberts,B., Johnson,A., Lewis,J., Raff,M., Roberts,K. and Walter,P. (2002) *Molecular Biology of the Cell*. 4th ed., Garland Publishing, London.

Arslan,A.N. *et al*. (2001) A new approach to sequence comparison: normalized sequence alignment. *Bioinformatics*, **17**, 327–337.

Benson,D.A. *et al*. (2004) GenBank: update. *Nucleic Acids Res*., **32**, 23–26.

Cormen,T.H., Stein,C., Rivest,R.L. and Leiserson,C.E. (2001) *Introduction to Algorithms*., Second Edition, MIT Press, Cambridge, MA, USA.

Heger,A. *et al*. (2004) Accurate detection of very sparse sequence motifs. *J. Comput. Biol*., **11**, 843–857.

Hopcroft,J.E. and Karp,R.M. (1973) An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM J. Comput*., **2**, 225–231.

Karolchik,D. *et al*. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*., **31**, 51–54.

Kellis,M. *et al*. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

Miller,W. (2001) Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics*, **17**, 391–397.

Morgenstern,B. *et al*. (1998) DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics*, **14**, 290–294.

Mouse Genome Sequencing Consortium (2002), Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Murphy,W.J. *et al*. (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science*, **294**, 2348–2351.

Notredame,C. *et al*. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol*., **302**, 205–217.

Park,J. *et al*. (1997) Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol*., **273**, 349–354.

Schwartz,S. *et al*. (2003) Human-mouse alignments with BLASTZ. *Genome Res*., **13**, 103–107.

Szklarczyk,R. and Heringa,J. (2004) Tracking repeats using significance and transitivity. *Bioinformatics*, **20** (Suppl. 1), I311–I317.

Thomas,J.W. *et al*. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, **424**, 788–793.

Vingron,M. and Waterman,M.S. (1994) Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J. Mol. Biol*., **235**, 1–12.

Ye,L. and Huang,X. (2005) MAP2: multiple alignment of syntenic genomic sequences. *Nucleic Acids Res*., **33**, 162–170.

Zhang,Z. *et al*. (1999) Post-processing long pairwise alignments. *Bioinformatics*, **15**, 1012–1019.